文章编号：1004-4353(2018)02-0143-06

基于融合特征网和模块网的低频行为挖掘方法

郝惠晶， 王丽丽， 刘祥伟
（安徽理工大学 数学与大数据学院，安徽 淮南 232001）

摘要：针对流程挖掘过程中忽略低频行为的问题，提出一种基于融合特征网和模块网挖掘低频行为的方法. 首先，通过处理有效的事件日志确定通讯行为轮廓关系，并根据日志将特征分为不同模块，重构事件内部行为，挖掘相应的模块网与特征网；然后，融合特征网与模块网得出完整的流程模型，并通过迭代扩展初始模式得出所有低频模式. 实例分析证明，本文提出的方法具有一定的可行性.

关键词：过程挖掘；特征网；模块网；低频行为
中图分类号：TP391.9 **文献标识码**：A

Low frequency behavior mining method based on feature nets and module nets

HAO Huijing， WANG Lili， LIU Xiangwei
（*College of Mathematics and Big Data*, *Anhui University of Science and Technology*, *Huainan* 232001, *China*）

Abstract: Aiming at the problem that the low frequency behavior is ignored in the process mining process, a method of mining low frequency behavior pattern based on fusion feature network and modular network is proposed in this paper. First of all, by processing effective event log to determine the communication behavior profile, dividing the characteristics into different modules according to the log, reconstructing the internal behavior of the event, thus, the corresponding module network and the feature network are excavated. Then, by integrating the feature network and the modular network, the complete process model is obtained, and iterating the initial mode, all the low frequency models is extended. Examples are given to prove the feasibility of the method.

Keywords: process mining; feature net; module net; low-frequency behavior

0 引言

随着大数据的不断发展，业务流程管理在很多领域发挥着重要作用. 过程挖掘作为业务流程分析的主要内容，大多为研究发现频繁行为. 但在现实中，一些低频行为虽然发生频数较低，但也是业务流程管理的核心内容之一，不可忽略[1]. 目前，已有很多学者对低频行为进行了研究，并取得了一些研究成果. 例如：文献[2]提出了一种利用整数线性规划的过程挖掘方法来发现合理的工作流网，并且设计了一种过滤算法来处理低频行为的存在. 文献[3]提出了一种挖掘局部流程模型的方法，并提出了一个增量程序来构建捕捉基于流程树的频繁模式的局部流程模型. 文献[4]通过 Inductive Miner 方法，利用切过滤不频繁行为，挖掘得到了一种合理的流程模型. 文献[5]提出了一个从日志中移除低频行为的自动技术，

收稿日期：2018-03-24 **作者简介**：郝惠晶(1993—)，女，硕士研究生，研究方向为 Petri 网.
基金项目：国家自然科学基金资助项目(61572035,61402011)；安徽省高校自然科学基金资助重点项目(KJ2016A208)；安徽理工大学研究生创新基金资助项目(2017CX2113)

该技术能够自动识别所有的行为并移除低频行为,并且将不能映射到清理后的行为识别为异常行为;但该技术存在一定的噪音,使得挖掘流程模型变得杂乱,容易出现异常值. 文献[6-8]提出了基于频率的不同噪音过滤方法,如机器学习技术、概率模型和专用噪音过滤方法,这些方法针对流程挖掘过程中出现的无序事件、异常行为等对模型具有影响的噪音进行了有效过滤. 文献[9]提出了一种从流程模型中检索低频行为模式的 Wo Mine-i 算法,该方法可以在过程模型中搜索选择、并发、循环的序列结构,但这些结构在日志中很少被执行. 上述文献对低频行为的研究都是要求已知完整的流程模型,对只知道系统运行所记录的事件日志并不适用. 针对这一缺陷,本文提出一种基于融合模块网和特征网的低频行为模式的挖掘方法,并通过实例分析证明本文方法的可靠性和有效性.

1 基本概念

定义 1[10](开放 Petri 网) $OPN=(P,I,O,T,F,i,f)$ 为开放 Petri 网,当且仅当:

1) $(P\cup I\cup O,T,F)$ 是一个 Petri 网.

2) P 是内部库所集,T 是变迁集,F 是库所和变迁之间弧线的集合.

3) I 是输入库所集,且 $\cdot I=\varnothing$;O 是输出库所集,且 $O^{\cdot}=\varnothing$.

4) P,I,O 两两不相交.

5) i 是初始标识,f 是终止标识.

一个 OPN 是封闭的,如果 $I=O=\varnothing$. $I\cup O$ 为 OPN 的接口库所.

定义 2[11](通讯行为轮廓) 设 $L\subseteq T^{*}$ 是事件日志,$\prec_L\subseteq T\times T$ 是相应的通讯后继关系. 通讯行为轮廓是一个三元数组 $(\rightarrow_C,\|_C,+_C)^{\mathrm{Com}}$,它由以下关系组成:

1) 严格通讯关系 $A\rightarrow_C B$,当且仅当 $A\prec_L B$,$B\not\prec_L A$;

2) 交叉通讯关系 $A\|_C B$,当且仅当 $A\prec_L B$,$B\prec_L A$;

3) 排它通讯关系 $A+_C B$,当且仅当 $A\not\prec_L B$,$B\not\prec_L A$;

4) 逆严格通讯关系 $A\leftarrow_C B$,当且仅当 $A\not\prec_L B$,$B\prec_L A$.

定义 3[12](特征网) 设 $L\subseteq T^{*}$ 是一个事件日志,$A,F\in T$ 是特征. 设 $(\rightarrow_C,\|_C,+_C)^{\mathrm{Com}}$ 是通讯行为轮廓,特征网 N_F 满足以下条件:

1) $P=\overline{P}$, $T=\overline{T}$, $i=[\overline{i}]$, $\Omega=\{[\overline{f}]\}$;

2) $I=\{p_{A-F}\mid A\rightarrow F\}$;

3) $O=\{p_{F-A}\mid F\rightarrow A\}$;

4) $F=\overline{F}\cup\{(t,p_{F-A})\mid t\in T,\lambda(t)=A,F\rightarrow A\}\cup\{(p_{F-A},t)\mid t\in T,\lambda(t)=A,A\rightarrow F\}$.

其中 I 和 O 是端口库所,$\langle\overline{P},\overline{T},\overline{F},\overline{i},\overline{f}\rangle$ 是工作流网.

定义 4[9](模式) 设 $C=(A,a_i,a_o,D,I,O)$ 是过程模型 M 的一个 C-网,A 的连通子图为 $P=(A',A'_i,A'_o,D',I',O')$. $A'_i\subseteq A'$ 和 $A'_o\subseteq A'$ 分别表示活动开始和活动结束,P 是 M 的一个模式,当且仅当 $A'\subseteq A$. $D'\subseteq D$,对任意的 $\alpha\in A'$:$I'(\alpha)\subseteq I(\alpha)$,$O'(\alpha)\subseteq O(\alpha)$. 如果一个模式 $P=(A',A'_i,A'_o,D',I',O')$ 是一个简单模式,当且仅当对于所有活动 $\alpha\in A'$:

$$[\exists!\ \Phi\in I'(\alpha):\Phi\not\subset R^{+}_{\alpha}]\vee[\forall\Phi\in I'(\alpha):\Phi\subseteq R^{+}_{\alpha}];$$

$$[\exists!\ \Theta\in O'(\alpha):\Theta\not\subset R^{-}_{\alpha}]\vee[\forall\Theta\in O'(\alpha):\Theta\subseteq R^{-}_{\alpha}].$$

其中 R^{+}_{α} 是活动 α 的所有后集集合,R^{-}_{α} 是活动 α 的所有前集集合.

定义 5[9](最小模式,M-模式) 给定过程模型 M 的一个 C-网 $C=(A,a_i,a_o,D,I,O)$ 和一个活动 $\alpha'\in A$,如果一个模式 $P=(A',A'_i,A'_o,D',I',O')$ 是 α' 的最小模式,当且仅当一个最大的简单模式包含 α',并且满足以下规则:

1) 如果 $|I(\alpha')|>1$，那么 $[I'(\alpha')=\varnothing] \vee [|I'(\alpha')|=1.\Phi \in I'(\alpha'):|\Phi|>1]$.

2) 如果 $|O(\alpha')|>1$，那么 $[O'(\alpha')=\varnothing] \vee [|O'(\alpha')|=1.\Theta \in O'(\alpha'):|\Theta|>1]$.

3) $\forall \alpha \in R_{\alpha'}^{+}$，如果 $|O(\alpha)| \neq 1$，那么 $O'(\alpha)=\varnothing$.

4) $\forall \alpha \in R_{\alpha'}^{-}$，如果 $|I(\alpha)| \neq 1$，那么 $I'(\alpha)=\varnothing$.

5) $\forall \alpha \in A', \alpha \neq \alpha', \alpha \in (R_{\alpha'}^{+} \cup R_{\alpha'}^{-})$，如果 $|I(\alpha)| \neq 1$，那么 $I'(\alpha)=\varnothing$，并且如果 $|O(\alpha)| \neq 1$，那么 $O'(\alpha)=\varnothing$.

2　基于融合模块网和特征网的低频行为挖掘分析

本文提出的基于融合特征网和模块网的低频行为模式的挖掘算法，主要是基于有效的事件日志确定系统的通讯行为轮廓关系，通过分析和融合相应的模块网和特征网，挖掘出一个完整的流程模型，基于流程模型测量其模式频率并与阈值比较发现其低频行为模式. 为判断频繁行为与低频行为，首先给出低频模式的定义.

定义 6[8]（低频模式）　设 L 是过程日志的迹集，一个简单模式 SP 的频率为 $freq(SP)=\frac{|\{\tau \in L: SP \vdash \tau\}|}{|L|}$，而模式 P 的频率为简单模式的最大频率，即 $freq(P)=\max\limits_{\forall SP \in P} freq(SP)$. 给定一个频率阈值 $\sigma \in \mathbf{R}: 0<\sigma \leqslant 1$，模式 P 是一个低频模式当且仅当 $freq(P)<\sigma$.

基于融合特征网和模块网的低频行为挖掘算法如下：

输入：事件日志，频率阈值 p

输出：低频行为

Step 1　整理事件日志 $L=\{l_1,l_2,\cdots,l_n\}$，对有效事件日志进行预处理，基于事件日志中各个活动事件之间的行为关系，建立通讯行为轮廓关系表.

Step 2　根据通讯行为轮廓关系，将系统分解为不同的模块，并挖掘出相应的模块网 $M_1,M_2,\cdots$，然后利用前驱后继关系建立特征网 M_F.

Step 3　挖掘接口库所 $I=\{p_{A-F} \mid A \rightarrow F\}$，$O=\{p_{F-A} \mid F \rightarrow A\}$ 和接口库所与特征之间的流关系 $F=\overline{F} \cup \{(t,p_{F-A}) \mid t \in T, \lambda(t)=A, F \rightarrow A\} \cup \{(p_{F-A},t) \mid t \in T, \lambda(t)=A, A \rightarrow F\}$，将模块网 M_1，$M_2,\cdots$ 和特征网 M_F 融合，挖掘出完整的流程模型.

Step 4　将事件日志的实例数从大到小进行排列，计算事件日志的频率 $freq$，如果 $freq>p$，则删除，构建剩余日志的流程模型. 初始化流程模型中活动的候选弧 $A^{<}$ 和 M-模式，开始扩展迭代，计算新模式的频率. 如果 $freq'>freq$，则舍弃，否则根据频率阈值 p 判断模式是否频繁；如果 $freq'>p$，则模式是频繁模式，需继续进行下一次迭代，否则执行 Step 5.

Step 5　如果 $freq'<p$，且满足以下其中一个条件，则迭代完成，输出流程模型的低频模式，算法结束：

i) 当前迭代为第 1 次迭代；

ii) $freq>p$，$freq'<p$；

iii) $freq=freq'$.

3　实例分析

实例分析以网上购物系统为例. 给定频率阈值为 0.1，考虑系统运行所记录的事件日志，共选择了 5 000条实例，如表 1 所示. 表 1 记录了用户系统(X)、商家系统(Y)、支付系统(Z)及 3 个系统交互的事件日志序列及其实例数，其中 A，…，W 分别表示登录、选择商品、加入购物车、满足优惠条件、使用优惠券、不满足优惠条件、支付、订单完成、支付失败、交易关闭、支付成功、反馈商品信息、库存足够、库存不

足、查询库存、商品优惠活动、验证支付信息、确认订单、备货、发货、反馈订单信息、确认支付、反馈支付信息. 首先对事件日志进行预处理,并建立通讯行为轮廓表(表 2).

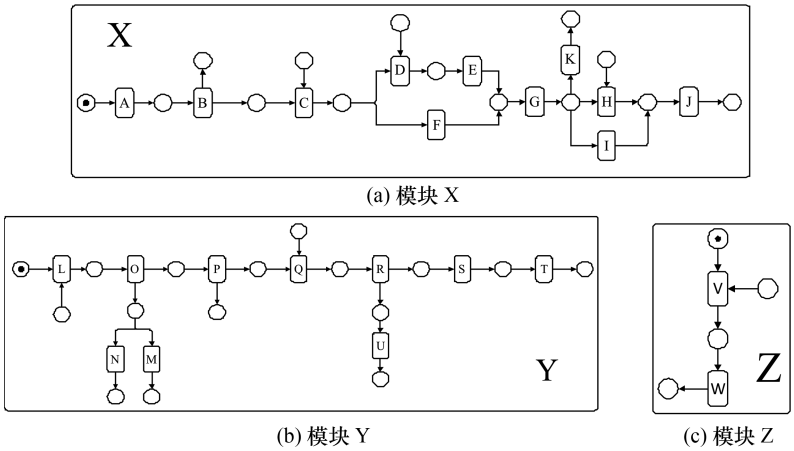
表 1 网上购物系统的事件日志

系统	事件日志	实例数	事件日志	实例数
X	ABCDEGHJ	1 304	ABCFGHJ	1 239
	ABCDEGIJ	45	ABCFGIJ	78
	ABCDEGK	1 151	ABCFGK	1 183
Y	LOPQRST	2 023	LON	89
	LOP	2 672	LOPQRU	2 014
	LOM	2 283		
Z	VW	2 895		
交互系统	LOPDE	2 356	PQRUH	2 014
	ABLOP	2 246	DEGKV	2 478
	LOMC	2 689	VWQR	2 344
	LONBC	89		

表 2 通讯行为轮廓关系表

	用户系统											商家系统										支付系统	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
A												+	+	+	+	+	+	+	+	+	+	+	+
B												→	+	←	+	+	+	+	+	+	+	+	+
C												←	←	+	+	‖	+	+	+	+	+	+	+
D												+	+	+	+	←	‖	‖		+	‖	+	+
E												+	+	+	+	+	‖	‖	+	+	‖	+	+
F												+	+	+	+	+	+	+	+	+	+	+	+
G												+	+	+	+	+	‖	‖	+	+	‖	+	+
H												+	+	+	+	+	+	+	+	+	←	+	+
I												+	+	+	+	+	+	+	+	+	+	+	+
J												+	+	+	+	+	+	+	+	+	+	+	+
K												+	+	+	+	+	+	+	+	+	+	→	+
L	+	←	→	+	+	+	+	+	+	+	+											+	+
M	+	+	→	+	+	+	+	+	+	+	+											+	+
N	+	→	+	+	+	+	+	+	+	+	+											+	+
O	+	+	+	+	+	+	+	+	+	+	+											+	+
P	+	+	+	→	+	+	+	+	+	+	+											+	+
Q	+	+	+	‖	‖	+	‖	+	+	+	+											+	←
R	+	+	+	‖	‖	+	‖	+	+	+	+											+	+
S	+	+	+	+	+	+	+	+	+	+	+											+	+
T	+	+	+	+	+	+	+	+	+	+	+											+	+
U	+	+	+	‖	‖	+	‖	→	+	+	+											+	+
V	+	+	+	+	+	+	+	+	+	+	←	+	+	+	+	+	+	+	+	+	+		
W	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	→	+	+	+	+		

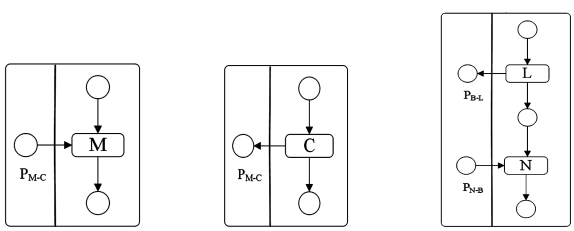
根据给定的事件日志及通讯行为轮廓关系,挖掘出相对应系统的模块网,如图 1 所示. 根据特征之间的通讯行为轮廓,可以发现事件的特征之间存在交互行为,并且有的特征既能接收信息也能发送信息. 重构每个事件的内部行为,找出对应的特征网,如图 2 所示. 将特征网和模块网融合得到完整的流程模型,如图 3 所示.



(a) 模块 X

(b) 模块 Y　　(c) 模块 Z

图 1　对应系统的模块网



特征C　　特征M　　特征B

图 2　事件的特征网

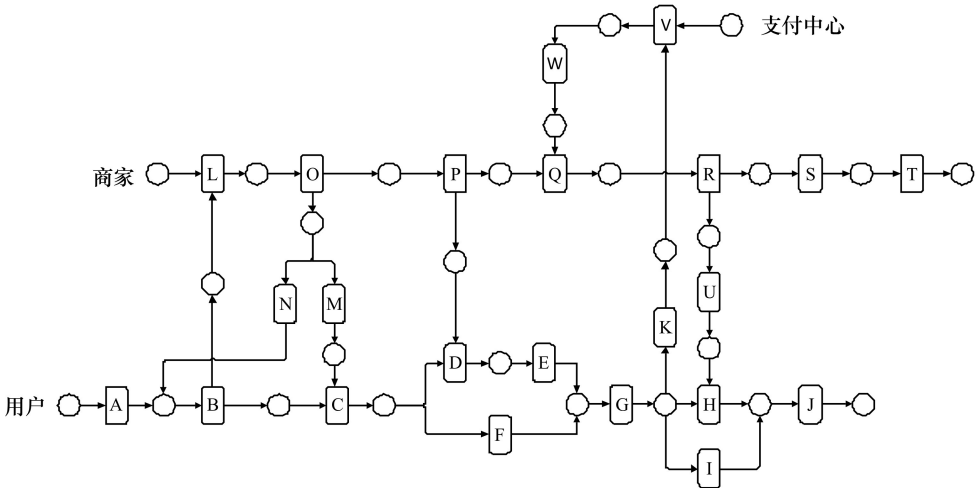


图 3　网上购物的完整流程模型

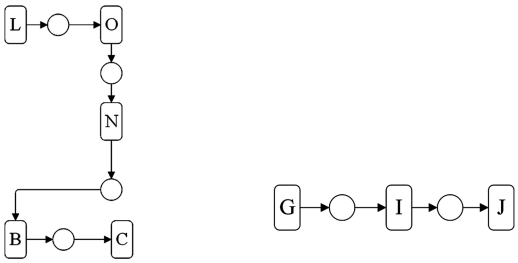
对事件日志从大到小排列如下：

$X=[\langle \mathrm{ABCDEGHJ}\rangle^{1304},\langle \mathrm{ABCFGHJ}\rangle^{1239},\langle \mathrm{ABCDEGK}\rangle^{1151},\langle \mathrm{ABCFGK}\rangle^{1183},\langle \mathrm{ABCFGIJ}\rangle^{78},$
$\langle \mathrm{ABCDEGIJ}\rangle^{45}]$；

$Y=[\langle \mathrm{LOP}\rangle^{2672},\langle \mathrm{LOM}\rangle^{2283},\langle \mathrm{LOPQRST}\rangle^{2023},\langle \mathrm{LOPQRU}\rangle^{2014},\langle \mathrm{LON}\rangle^{89}]$；

$Z=[\langle \mathrm{VW}\rangle^{2895}]$.

为了发现流程模型中的低频行为模式，需先将频数较高的事件日志删除. 根据上述 3 个系统的事件日志实例数的排列情况，将每个系统频数较高的日志删除，对剩余日志中活动的 M-模式进行迭代扩展. 经计算得活动 N 和活动 I 的 M-模式(图 4) 出现的频数较低，再由上述算法得这两个 M-模式在第 1 次迭代结束后即为低频模式，$f(N)=0.0178<0.1$，$f(I)=0.0246<0.1$. 其余模式因频率大于给定阈值，故直接删除.



(a) N的M-模式 (b) I的M-模式

图 4 低频模式

4 结论

本文给出的基于融合特征网和模块网的低频行为模式的挖掘方法,不仅能够找出简单事件日志中的低频行为,而且利用迭代扩展的方式大大缩小了人工计算量,具有一定的应用价值.在挖掘的低频行为中,如何有效地区分低频和噪音,以及如何过滤噪音将是我们今后的研究方向.

参考文献:

[1] Sani M F, van Zelst S J, van der Aalst W M P. Improving process discovery results by filtering outliers using conditional behavioural probabilities[C]//International Conference on Business Process Management. Cham: Springer, 2017:216-229.

[2] van Zelst S J, van Dongen B F, van der Aalst W M P, et al. Discovering relaxed sound workflow nets using integer linear programming[J]. Computing, 2018,100(5):529-556.

[3] Tax N, Sidorova N, Haakma R, et al. Mining local process models[J]. Journal of Innovation in Digital Ecosystems, 2016,3(2):183-196.

[4] Leemans S J J, Fahland D, van der Aalst W M P. Discovering Block-Structured process models from event logs containing infrequent behaviour[C]//International Conference on Business Process Management. 2014,171:66-78.

[5] Conforti R, Rosa M L, Hofstede A H M T. Filtering out infrequent behavior from business process event logs[J]. IEEE Transactions on Knowledge & Data Engineering, 2017,29(2):300-314.

[6] Liesaputra V, Yongchareon S, Chaisiri S. Efficient process model discovery using maximal pattern mining[C]//International Conference on Business Process Management. Cham: Springer, 2015:441-456.

[7] Bellodi E, Riguzzi F, Lamma E. Statistical relational learning for workflow mining[J]. Intelligent Data Analysis, 2016,20(3):515-541.

[8] Carmona J, Broucke S K. Incorporating negative information in process discovery[C]//International Conference on Business Process Management. New York, Inc: Springer-Verlag, 2015:126-143.

[9] Chapela-Campa D, Mucientes M, Lama M. Discovering infrequent behavioral patterns in process models[C]//International Conference on Business Process Management. Cham: Springer, 2017:324-340.

[10] 程腾腾,方贤文,王丽丽,等.融合特征网与模块网的业务过程挖掘[J].计算机工程与应用,2017,53(20):237-242.

[11] van der Werf J M, Kaats E. Discovery of functional architectures from event logs[C]//PNSE@ Petri Nets. 2015: 227-243.

[12] 谢苗苗,刘祥伟,王丽丽.基于通信行为轮廓的手机充值流程挖掘方法[J].牡丹江师范学院学报(自然科学版),2017(4):11-15.